Proposition de stage

Entreprise : IPEDIS

Titre du sujet de stage : Analyse et traitement automatique des documents pdfs

Encadrant : Christian KAHINDO (IPEDIS)

Co-encadrante : Nicole VINCENT (LIPADE, Université de Paris),

Durée du stage : 5 mois

Rémunération : 900 euros brut / prime de 1000 euros à la fin du stage en fonction des résultats

Mots-clés : Documents pdfs, perception visuelle, modélisation, intelligence artificielle symbolique, argumentation computationnelle, SDK Adobe Acrobat, contenu pdfs.

Candidatures : Elles doivent être envoyée à joinus@ipedis.com, joindre un CV et le relevé de notes de M1. Il faut mentionner dans l'objet du mail « stage R&D A11y ».

Ce stage s'inscrit dans le cadre d'un projet mené par la Société IPEDIS et auquel a participé le laboratoire d'informatique (LIPADE) de l'Université de Paris. Ce projet a pour objectif de développer un système de balisage automatique qui permettra de produire un document pdf accessible à partir d'un document pdf non-accessible.

Le but de ce stage est d'analyser la mise en page des documents pdfs à partir de l'extraction automatique de ses paragraphes en fonction d'une modélisation basée sur la perception visuelle. Cette étude basée sur les principes du traitement d'image et sur le traitement automatique du langage naturel permettra d'identifier, de détecter et d'analyser par exemple la table des matières, les notes et légendes [1] contenues dans un document en s'inspirant des méthodes exploitées sur les images de documents. On pourra aussi identifier les références, les tableaux matérialisés/semi-matérialisés et non-matérialisés [2], les citations, la présence d'hyperliens, les césures de mots, etc.

Dans ce contexte on partira des acquis suivants :

- La manipulation des éléments constitutifs d'un fichier pdf ;
- La manipulation et le balisage du contenu des pdfs à partir du SDK d'Adobe Acrobat ;
- La segmentation du texte en paragraphes et la caractérisation de ces derniers comme titre ou paragraphe standard par exemple. Cette dernière repose sur des méthodes fondées sur le cadre théorique d'argumentation computationnelle appelé « Logic Programming with Priorities (LPP) » [3] et la méthodologie de construction de théories argumentatives appelée « Scenario-Based Preferences » [4].

Mission du stage (détail des activités) :

- Une étude bibliographique sur le sujet sera demandée (extraction des légendes, tableaux, notes, etc.) ;

- Analyser par quels éléments du contenu des fichiers pdf sont constituées les structures telles qu'une table des matières, une légende, une référence, une citation, un tableau, etc. Les positions des différents éléments devront également être prises en compte.

- Définir et implémenter une solution permettant d'identifier et d'interpréter les différents éléments de constitution du document.

Profil du candidat

Etudiant en dernière année cycle d'ingénieur ou en Master 2, vous avez suivi une formation en traitement du signal et des images ou en intelligence artificielle.

Bibliographie

[1] Khurshid K., Faure, C., Vincent, N. ***Fusion of Word Spotting and Spatial Information for Figure Caption Retrieval in Historical Document Image***. In Conference on Document Analysis and Recognition (ICDAR), Barcelona, Spain, pages 266-270, 2009.

[2] Alhéritière Hélo"ise; Ama"ieur Walid; Cloppet Florence; Kurtz Camille; Ogier Jean-Marc and Vincent Nicole ***Straight Line Reconstruction for Fully Materialized Table Extraction in Degraded Document Images***. In Discrete Geometry for Computer Imagery - 21st IAPR International Conference, DGCI 2019, Marne-la-Vallée, France, March 26-28, 2019, Proceedings, pages 317-329, 2019.

[3] Kakas A., Moraitis P., *"**Argumentation Based Decision Making for Autonomous Agents**", in Proc. 2nd International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'03), pp. 883-890, Melbourne, Australia, 2003.*

[4] Kakas A., Moraitis P., *Spanoudakis N., "**GORGIAS: Applying argumentation**", Argument & Computation, Vol. 10, No. 1, pp. 55-81, 2019.*